

Avsnitt 3. Bra och dåliga modeller. Testning

George Box: "Det finns inga sanna modeller, men några är nyttiga".

Vi går tillbaka till föregående kapitel och talar om datageneratorn, som är en ganska bra metafor på hur data blir till. Vi tänker oss det finns en mycket komplicerad mekanism, M som har genererat den eller de tidsserier y vi är intresserad av att förstå. Den registrerade verkligheten (tidsserien) är en realisation som kommer till oss så att någon matat in slumpantal e i denna modell och ut kom vår tidsserie. Generatorn får följande utssende:

$$y = M + e.$$

Här är alltså e i.i.d. $N(0, s^2)$, all systematik och allt som finns att förstå ingår i M . Vad vi försöker oss på är just att ta reda på hur detta M kunde se ut, att "förstå y ". Ett sätt att utnyttja denna insikt är att använda vårt sampel y för att göra ett bästa estimat på M , lå oss kalla det $\hat{M}$. Sedan kan vi generera observationer utanför den period från vilken våra observationer y kommer. Då gör vi en **prognos** med vår estimerade **modell**:

$$y = M + a.$$

Prognosen får vi genom att sätta värdena på a till sitt väntevärde 0, för vi vet ju inte vilka slumpantal som genererar framtida observationer. Är $\hat{M}$ en bra modell av M borde de genererade prognosvärdena vara mycket sannolika, beroende på hur stor variansen på a är.

Men för att nu $\hat{M}$ skall vara ett bra estimat av M måste vi se till att den innehåller all systematisk information om M som går att få ur observationerna y . För om det inte är så betyder det att en del

systematik/förklaring ligger kvar i a och då har vår dekomponering i "förstått" (M) och "inte-förstått" (a) inte lyckats; M är ingen modell och a är ingen residual. Vi är då tillbaka i ruta ett. Börja om från början.

Obs! En fara ligger i att man koncentrerar sig på modellen. Man vill ha stöd för någon förutfattad mening man har. Man får kanske signifikanta t-värden för koefficienterna i en regression. Så struntar man att kolla residualerna. Om de är autokorrelerade är det lätt att få höga t-värden, vilka är missvisande och beroende på bias i estimatet för felvariansen. Därför är det lika viktigt att se på residualerna som att se och tolka modellen. Det görs genom **testning**. Man talar ibland om **modelldiagnostik**.

Man kan säga att så länge en ekonom för ett resonemang inom ett teoretiskt system (axiomatik) räcker det med matematisk stringens. Såfort han/hon för ut resonemanget i verklighetens siffervärld är det statistiken som måste ta över för att stöda eller förkasta en hypotes. Vi skall gå igenom några sätt att undvika dåliga modeller:

Att försöka klämma mer ur data än där finns att ta

David Hendry: "If you torture your data enough it will confess".